Text Mining

Nascita ed evoluzione della statistica testuale

- Vengono sviluppati strumenti *software* e vengono proposte metodologie per l'elaborazione dei dati testuali.

- Mutano anche i soggetti stessi, protagonisti di questi studi: non solo università o centri di ricerca, ma anche aziende che operano on line and off line oppure solamente on line.

(Nel seguito si fa riferimento principale a Bolasco (2005))

Nascita ed evoluzione della statistica testuale

- Le aziende dovendo interagire con enormi masse di materiali testuali disponibili in rete (l'80% delle informazioni in azienda è in forma di testi e solo il 20% in dati numerici), hanno il problema di selezionare, all'interno di fonti smisurate, i dati di loro interesse, per estrarne *informazione* capace di produrre valore.
  - Si tratta di soluzioni di *Text Mining* orientate alla gestione della conoscenza e alla cosiddetta *Business Intelligence*.

Nascita ed evoluzione della statistica testuale

Primi esempi di analisi quantitativa applicata a dati testuali grazie a Zipf (1935), Yule (1944), Guirard (1954) e Herdan (1956).

Nel seguito Benzecri fonda i suoi studi sull'analisi di dati linguistici e formalizza, nel 1973, ***l'analyse des donnes***, seguendo la tesi dell'induttività linguistica opposta alle idee di linguisti precedenti.

Vengono poi sviluppati da Muller (1973) e Laffon (1984) i primi indici e misurazioni divenuti classici nella statistica linguistica e lessicale. Parallelamente, in Italia vengono creati i primi cosiddetti lessici di frequenza dai linguisti Zampolli (1971) e De Mauro (1980).

- Il focus si sposta dall'analisi di testi veri e propri (es. ambito letterario) allo studio di testi artificiali e in generale di dati espressi in linguaggio naturale raccolti da fonti disparate:

    - indagini sul campo (domande aperte o interviste);

    - analisi di frammenti o testi corti (abstract, bibliografie, manifesti, messaggi),

raccolti in una collezione di documenti costituente un *corpus di dati testuali*.

- Il corpus può essere studiato secondo la sua frammentazione in documenti o records.

- Un vantaggio dell'analisi automatica su base statistica consiste nell'essere indipendente dall'ampiezza o dimensione dei testi che hanno originato la raccolta.

- Alla fine degli anni ‘80, Benzecri, Lebart e Salem definiscono i confini della *statistica testuale* basata sull’analisi per *forme grafiche* (e non più per *lemmi*) ed in parallelo sviluppano *software* per l’analisi dei dati testuali:
  - *Addad*
  - *Spad_T* che fa impiego di metodi multidimensionali, come le analisi fattoriali su matrici sparse con calcolo degli autovalori in lettura diretta ;
  - *Lexico* che consente l’individuazione nel corpus dei *segmenti ripetuti* e l’analisi delle *specificità* per l’estrazione di parole caratteristiche delle sub-parti grazie ad un test basato sulla legge ipergeometrica.
  - *Alceste*

Le diverse unità di analisi del testo

- Il problema essenziale per un'analisi automatica di un testo è operare il riconoscimento del senso.
- Con il termine *parola* si indica convenzionalmente l'unità di analisi del testo. A seconda degli obiettivi, tale unità può essere

una forma grafica un'*unità mista* detta l*essia* ((semplice: <carta>; composta: <carta geografica>; complessa: <carta di credito>)),etc… in grado di catturare al meglio il contenuto presente nel testo.

- J.P. Benzécri (*Addad,* 1981), A. Salem (*Lexicloud,* 1987) e Reinert (*Alceste*, 1986-2003), con i loro software, mostrano che partendo da un'analisi formale si arriva a cogliere la struttura del *senso* presente nel corpus di testi.
- Da un'analisi di tipo *paradigmatico (vocaboli)*, in cui le parole sono listate in un qualche ordine (alfabetico, inverso…), si ottiene una rappresentazione della struttura *sintagmatica (senso)* presente nel testo/corpus.

- L’ambiguità insita nel linguaggio viene risolta attraverso l’analisi complessa di *grandi matrici di dati testuali* grazie ai metodi e alle tecniche di *analisi multidimensionale* (analisi delle corrispondenze, cluster analysis, analisi discriminante, multidimensional scaling).
- Tali analisi, misurando la *similarità di profili lessicali*, producono rappresentazioni contestuali dell’informazione testuale che si traducono in visualizzazioni nelle quali vale il principio della vicinanza *vs* somiglianza delle unità lessicali.

- Attraverso un’analisi fattoriale, ad esempio è possibile in alcuni casi ricostruire delle frasi modali. (Bolasco, 1999), utilizzabili come veri e propri *modelli di senso* del contenuto del testo.
- Un altro esempio di utilizzo di assi semantici latenti è quello utilizzato nell’approccio detto *semiometrico* (L. Lebart *et al.* 2003): a partire da un set di 200 parole-stimolo ad alto contenuto simbolico, è possibile posizionare un campione di intervistati secondo alcune dimensioni di senso ricostruibili stabilmente nelle culture occidentali, molto utili nelle analisi di marketing.

- Accanto a questa analisi statistica si sistematizza la formalizzazione linguistica di particolari classi di parole (ad esempio tavole dei verbi, di forme composte (avverbi, preposizioni e gruppi nominali) e si sviluppano strumenti concreti di *lessicografia e linguistica computazionali* quali dizionari elettronici e automi per la descrizione di grammatiche locali .

Nasce poi un nuovo approccio di tipo lessico-testuale, l'unità di analisi è più flessibile, ad esempio una **lessia** (semplice: <carta>; composta: <carta geografica>; complessa: <carta di credito>).

Previste le forme composte, quando sono la rappresentazione minimale del significato che si intende catturare in fase di parsing. Particolare attenzione ai gruppi nominali quali:

- Aggettivo_Nome (terzo mondo, estratto conto)
- Nome_Aggettivo (lavoro nero, carta bianca)
- Nome_Preposizione_Nome (ordine del giorno, chiavi in mano)

In generale i casi in cui l'espressione composta ha significato diverso dalla somma dei significati delle parole che la formano.

Analisi Automatica dei testi

Rappresentazione adeguata del corpus prevede 4 fasi:

1) Preparazione del testo;
2) Analisi lessicale;
3) Estrazione di informazione;
4) Analisi testuale.

1) **Preparazione:** pulizia e definizione del charset, normalizzazione (spazi, apostrofi, accenti, entità particolari), tagging con meta-informazioni (classe grammaticale, lemma di appartenenza, numero di occorrenze nel corpus).

Analisi Automatica dei testi

Analisi lessicale: tipo verticale, rappresentazione paradigmatica del corpus, studio del vocabolario e del linguaggio (bag of words), statistiche su parole “piene” (verbi, nomi, aggettivi) e “vuote” (incipit, congiunzioni, punteggiatura), individuazione costanti del linguaggio.

3) Estrazione di informazione: selezione linguaggio peculiare senza query attraverso risorse esogene (scarto d’uso della parola dalla frequenza d’uso di riferimento) o endogene. Con query attraverso indici appositi.

- TFIDF (*term frequency–inverse document frequency*)
- Indice IS per l'analisi delle sequenze

L'analisi automatica dei testi

TFIDF (*term frequency–inverse document frequency*)

$$w = tf \times \log\left(\frac{N}{n}\right)$$

- tf è la frequenza del termine richiesto in ciascun documento;
- n il numero di documenti contenenti quel termine;
- N il numero totale di documenti presenti nel corpus.

Questo indice pondera quindi le parole in funzione della loro rilevanza, ossia tanto più esse sono frequenti esclusivamente in pochi documenti.

L'analisi automatica dei testi

L’indice IS filtra i segmenti rilevanti secondo la loro capacità di assorbimento delle occorrenze delle parole componenti.

$$IS = (\sum_{i=1}^{L} \frac{f_{segm}}{f_{gi}})P$$

Questo indice somma i rapporti di composizione delle occorrenze delle L parole appartenenti al segmento richiesto, ponderando tale somma con il numero P di parole piene.

Analisi Automatica dei testi

4) **Analisi testuale:** operazioni rivolte direttamente sul corpus, rappresentazione sintagmatica del testo, analisi delle concordanze e delle co-occorrenze.

Applicazione di operazioni simili nelle fasi di analisi lessicale e di analisi testuale, alle unità di testo che costituiscono il vocabolario (*analisi lessicale*) e alle unità di contesto, che costituiscono l'intero corpus come insieme totale delle occorrenze (*analisi testuale*). Esempi: individuazione costanti del discorso, produttività delle parole, analisi delle concordanze, ricerca entità d'interesse…

Analisi lessicale e analisi testuale

- Operazioni dello stesso tipo possono applicarsi sia in analisi lessicale alle unità di testo (parole o lessie) costituenti il vocabolario (V), sia in
- analisi testuale alle unità di contesto (documenti o frammenti del discorso.) costituenti il corpus come insieme totale delle occorrenze (N).

Individuazione costanti del discorso: primo screening sui termini più frequenti, classi grammaticali con incidenza più alta rispetto al lessico comune.

- Esempio Bolasco (2005) studio sul linguaggio eno-gastronomico, in cui l'analisi degli aggettivi evidenzia un eccesso di qualificazione. Come è naturale aspettarsi in una Guida, vi è una marcata tendenza alla positività (*buono, ottimo, grande, bello* sono gli aggettivi più frequenti)

Quando una parola è molto frequente in un corpus è altamente probabile che la sua produttività morfologica in quel testo sia elevata.

Produttività delle parole: la capacità di una parola di generare forme derivate da una radice comune.

- Esempio Bolasco (2005) studio di dieci annate del quotidiano *La Repubblica*.

- Nel grafo si illustra il caso del lessema <politic_> che, nel corpus Rep90, produce una varietà di 198 forme grafiche diverse per un totale di 344.930 occorrenze. Di queste il 99,4% riguarda le quattro forme base (politica/o/i/he) e lo 0,6% le altre formazioni che, espresse in lemmi, si articolano in 128 prefissi (2.530 occ.) e/o 28 suffissi (1.869 occ).

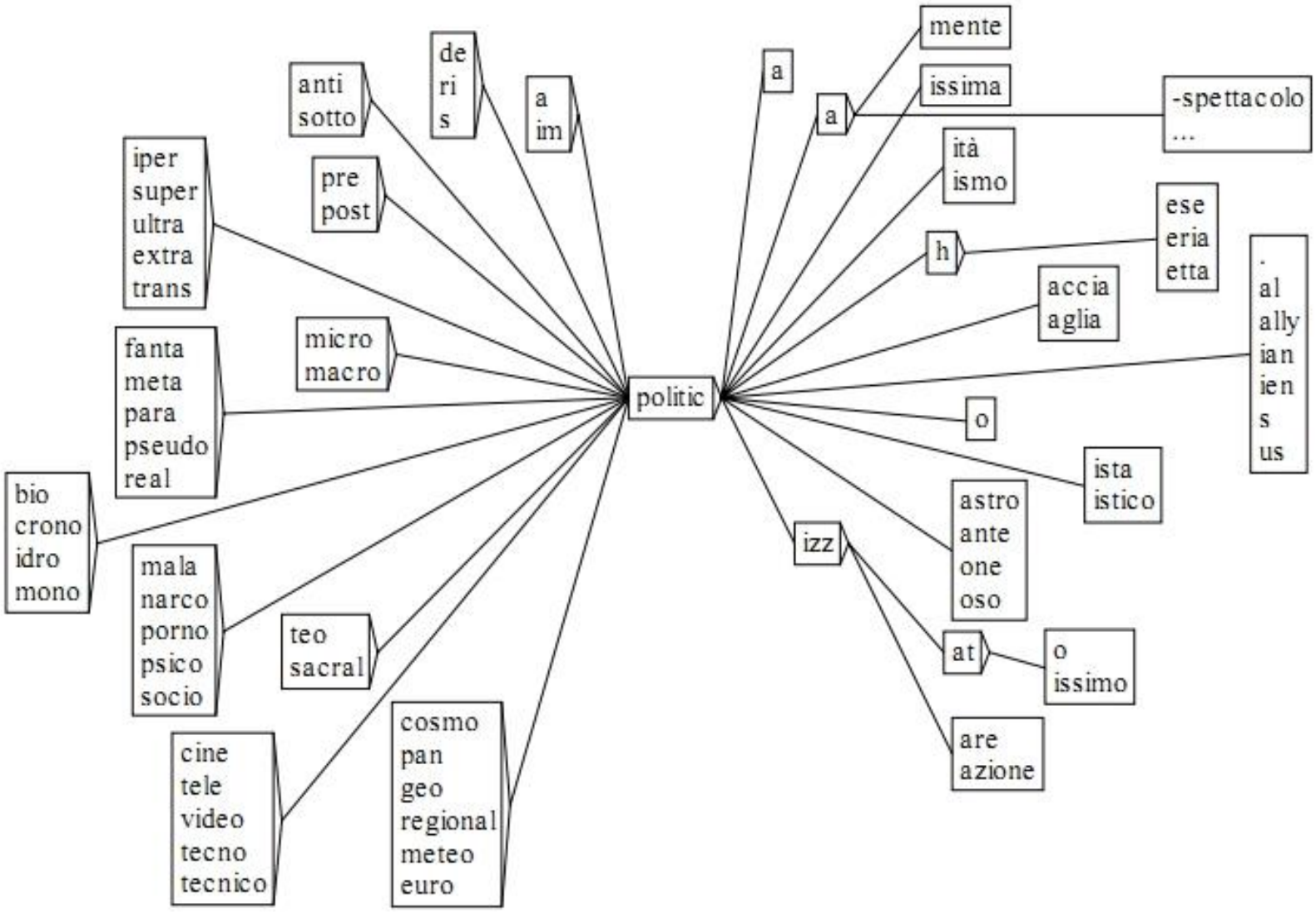


Figura 1. Grafo dei prefissi e suffissi della base <-politic-> in Rep90.

Analisi delle concordanze: fornisce l'insieme dei *co-testi* destro e sinistro di una parola pivot, tecnica utile per distinguere il significato reale di ogni occorrenza del vocabolo all'interno del corpus. Inoltre in questo modo si ricostruiscono per ogni parola i riferimenti tematici a cui questa rinvia, tracciando una mappa concettuale fra parole e temi affrontati.

- L’analisi delle concordanze fornisce l’insieme dei *co-testi* destro e sinistro di una predefinita parola *pivot* è utile per discernere il significato reale di ogni occorrenza di un vocabolo.
- La Tabella (Bolasco 2005) riporta esempi di *query* applicabili sia da unità di testo sia a unità di contesto (documenti del corpus). In quest’ultimo caso, il risultato produce l’estrazione dei documenti che le verificano, con la evidenziazione delle singole occorrenze in modalità *fulltext*.

Tabella 7 – Esempi di queries per concordanze su unità di testo (vocabolario) e unità di contesto (documenti del corpus).

A - Query sul vocab.	B - Query complessa sui frammenti del corpus (unità di contesto)			
Ricerca del lessema *auto**		Ricerca dei frammenti contenenti un elemento del concetto: *parentela*	flessioni	forme attualizzate
auto	1	padre/i OR madre/i OR mamma/e OR babbo/i	8	5
automobile	2	papà OR papa'	2	1
autobus	3	figlio/a/e/i OR figliola/o/e/i	8	6
autovettura	4	marito/i OR moglie OR mogli	4	3
autostrada	5	fratello/i OR sorella/e OR frat(sor)ellino/a/e/i	12	7
autocarro	6	suocero/a/i/e	4	3
autogrill	7	genitore/i	2	2
autodromo	8	nonno/a/i/e OR bisnonno/a/i/e	8	5
autosalone	9	nipote/i OR nipotino/a/i/e	6	6
autotreno	10	zio/a/i/e	4	3
autofficina	11	cognato/a/i/e OR cugino/a/i/e OR cuginetta/o/i/e	12	12
autolavaggio	12	genero/i OR nuora/e	4	2
automezzo	13	parente/i OR familiare/i OR famigliare/i	6	6
	totale flessioni e forme attualizzate		80	61

- *Ricerca di entità d'interesse*: Si vuole, ad esempio, estrarre da un corpus tutte le occorrenze relative all'entità <impresa>. Approccio di tipo dizionario, si incrociano i dati testuali a disposizione con un'anagrafica delle imprese.

- esempio Bolasco (2005) provvedimenti emessi dall'Antitrust sulle concentrazioni.
- Il corpus in questione (Baiocchi *et al.* 2005), di oltre 3500 provvedimenti, viene sottoposto preliminarmente al riconoscimento di tutte le imprese la cui ragione sociale è citata in maniera completa. Questo avviene grazie ad un dizionario di imprese contenente la ragione sociale. (forma giuridica inclusa) di ogni società,

Estrazione di informazione con risorse esogene o endogene: estrazione del linguaggio peculiare, ovvero quella parte di vocabolario utilizzato nel corpus che caratterizza maggiormente il testo. Avviene per comparazione, sulla base di un lessico di frequenza che rappresenti il linguaggio comune di una determinata comunità linguistica.

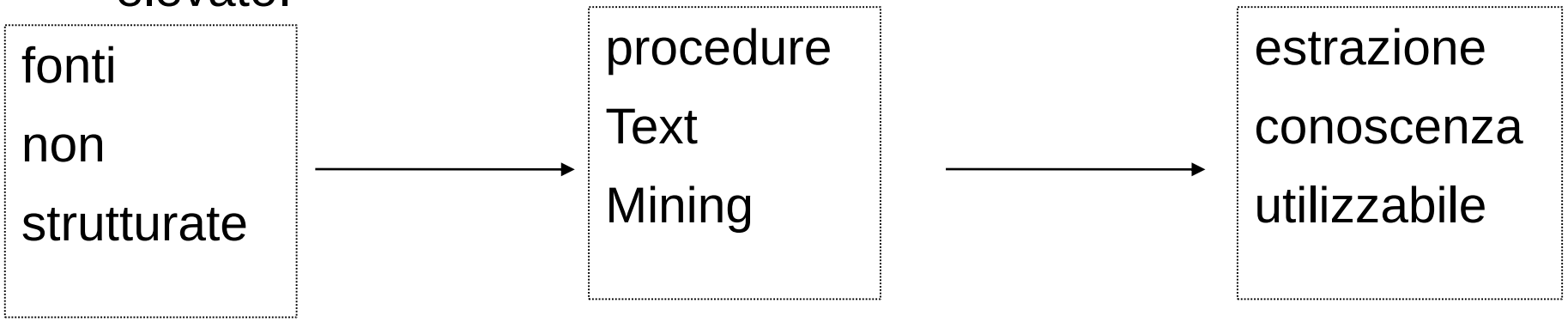
Avendo a disposizione dei dati storici, come ad esempio 10 anni di pubblicazioni del quotidiano di *Repubblica* (corpus *Rep90*, Bolasco 2005) è possibile estrarre ad esempio valutare il ciclo di vita delle parole calcolando gli scarti d'uso da un anno all'altro.

Text Mining

Disponibilità sempre maggiore di risorse informatiche permette di decuplicare la dimensione dei testi analizzabili ogni 2-3 anni.

Sviluppo tecniche di Text Mining per fronteggiare l'eccesso di informazione disponibile, procedure di *Information Retrieval* (reperimento delle informazioni) e *Information Extraction* soprattutto in ambito aziendale e istituzionale.

Differenze con Data Mining, dati testuali sono più espliciti ma più ostici per un'analisi automatica, potenziale commerciale elevato.



Text Mining

Presenza di un *document warehouse*, dove i testi sono raccolti, come corpus di dati testuali da analizzare.

Fornisce supporto alla comprensione del contenuto di un testo senza doverne leggere ogni parola.

Le applicazioni ricadono in diverse aree tra cui:

- Esplorazione dei dati testuali e del loro contenuto
- Uso di queste informazioni per migliorare processi esistenti

Nel primo caso intenti descrittivi, con lo scopo di estrarre temi e concetti presenti in un testo (es. *Customer Satisfaction*).

Nel secondo caso mining predittivo, classificazione di nuovi documenti in categorie definite con un *training set* di partenza.

Text Mining

Text Mining come incontro virtuale fra tecnologie informatiche, linguistiche e statistiche. Sono previste tre fasi:

- **Pre-processing**, prevale l'aspetto informatico, reperimento e formattazione (*es. XML*) dei testi e costruzione del document warehouse
- **Parsing del testo**, prevale l'aspetto linguistico, riconoscimento dei vocaboli e loro lemmatizzazione (divisione sostantivi, aggettivi e verbi)
- **Analisi dei documenti**, prevale l'aspetto statistico, categorizzazione automatica di documenti secondo diversi criteri: identificazione tematiche, relazioni o *supervised* in categorie predefinite. Clusterizzazione *unsupervised* basata sulla similarità del vocabolario senza categorie predefinite.

Text Mining

3 macro-aree:

Estrazione di informazioni per il "consumo umano": produrre riassunti, estratti e abstract in modo automatico; problemi di document retrieval con input di query.

Stima della similarità di documenti: categorizzazione supervised e unsupervised; in particolare identificazione del linguaggio in corpora multilingua e deduzione dell'autore dall'analisi del contenuto del testo (approccio *stilometrico*).

Estrazione di informazioni strutturate: ricerca di entità particolari (nomi persone, società, luoghi,..), studio relazioni fra entità, estrazione di regole dal testo.

Text Mining

Mining di testi strutturati.

Molti dei testi digitalizzati, ad esempio reperiti in rete, contengono una marcatura esplicita (es. HTML) e pertanto sono diversi dal testo semplice per quanto riguarda un approccio quantitativo.

In particolare si distinguono marcature interne, che descrivono informazione sulla struttura e sul formato del testo, e marcature esterne e forniscono informazioni sui collegamenti ipertestuali fra documenti diversi.

Entrambe le tipologie di informazioni possono essere sfruttate per migliorare le procedure di TM.

Campi applicativi del TM.

- *Customer Relationship Management* (CRM): classificazione e indirizzamento automatico delle e-mail, mediante integrazione di tecnologie statistiche di classificazione (basate su parole chiave e/o su concetti) e tecnologie linguistiche di estrazione della informazione, basate sulla comprensione del testo contenuto nel messaggio.
- *Customer Opinion Survey*: analisi automatica delle segnalazioni e/o reclami pervenuti per telefono o posta elettronica; monitoraggio costante delle opinioni espresse dai clienti in forum di discussione virtuale, come newsgroup e chat; analisi di domande aperte nelle survey quali/quantitative.

- *Gestione delle risorse umane*: controllo della motivazione aziendale a partire dall'analisi automatica delle opinioni espresse dai dipendenti in occasione di apposite rilevazioni; analisi dei curriculum vitae on-line per l'estrazione di specifici skills professionali.
- *Osservazioni sulla concorrenza e sull'utenza*: monitoraggio della situazione del mercato sia in termini di potenziali clienti che di concorrenti mediante il reperimento sul Web di liste di aziende, corredate dalle informazioni desiderate; analisi dell'immagine aziendale così come emerge dall'esame automatico di notizie e articoli.

- *Technology Watch* e analisi dei brevetti: ricerca e archiviazione sistematica di informazioni sulle tecnologie esistenti per l'identificazione dei settori in maggiore sviluppo; analisi automatica delle informazioni testuali contenute nei brevetti per identificare settori di ricerca emergenti.
- *Analisi di basi documentali settoriali* (economico-finanziarie, giuridiche, epidemiologiche, medico-farmaceutiche ecc.) con estrazione automatica di contenuti, riconoscimento di argomenti e relativa categorizzazione semantica.

- *Natural Language Processing*: costruzione di risorse linguistiche e di basi di conoscenza specifiche (dizionari, grammatiche, reti semantiche) e predisposizione di sistemi per la gestione di interrogazioni in linguaggio naturale, ad esempio nell'ambito di sistemi di *e-government*.
- Anche nelle attività di *Intelligence* riguardanti problemi di sicurezza nazionale è sempre più diffuso l'utilizzo di tecnologie di TM. In particolare, ad esempio nelle analisi multilinguistiche di vasti giacimenti di informazioni sul web e nell'identificazione del l' autore del testo).

I *settori* maggiormente interessati dal TM sono

- Dell'editoria e dei media (archivi multimediali automatizzati di grandi gruppi editoriali);
- delle telecomunicazioni, energia e altre aziende di servizi (call-center, portali web per servizi alle piccole e medie imprese);
- Dell'Information Technology e Internet (NLP, risorse linguistiche *online*, traduttori automatici);

- delle banche, assicurazioni e mercati finanziari (CRM, analisi del rischio finanziario e della comunicazione finanziaria d.impresa);
- delle istituzioni politiche, della Pubblica Amministrazione e della documentazione giuridica (analisi documentale, informazione istituzionale *on-line*, interrogazioni in linguaggio naturale);
- del settore farmaceutico e sanitario (estrazione automatica dei dati da abstracts a contenuto biomedico, gestione dei dati clinici).

Applicazione

Cluster Analysis sui dati forniti dal Prof. Palumbo. 3 testi sottoposti a team di 5 traduttori della SSLMIT.

Scopo: confermare l'ipotesi che tanto lo stile di scrittura quanto l'argomento di un testo possono essere fonti di difficoltà nella traduzione, riferimento particolare alla presenza di *grammatical metaphor*.

Metafore lessicali (uso parole con diverso significato), metafore grammaticali (variazioni nell'espressione di uno stesso significato).

Lo stesso significato può essere espresso secondo realizzazioni cosiddette *congruenti* o *incongruenti*, a secondo della presenza o meno di metafore grammaticali.

Applicazione

Esempio:

Congruent: *the driver drove the bus too rapidly down the hill so the brakes failed.*

Incongruent: *the driver's overrapid downhill driving caused brake failure.*

Due espressioni con lo stesso significato, la prima ha una costruzione *unmarked* (senza metafore), la seconda ha una costruzione *marked*, due verbi (drove, failed) trasformati in sostantivi (driving, failure).

Tendenza alla nominalizzazione come tratto più distintivo di una metafora grammaticale.

Testi da sottoporre selezionati dunque in base al livello di metaforicità, calcolato secondo opportuni indici linguistici, così disposti: ST3 < ST1 < ST2.

Applicazione

Selezione soggetti partecipanti motivata puramente da ragioni opportunistiche, a seconda della disponibilità data dagli stessi.

Stesso gruppo di 5 traduttori per tutto il percorso investigativo.

Gruppo omogeneo sotto due profili: completa mancanza di familiarità con l'argomento trattato nei testi e capacità di maneggiare testi scientifici sia in inglese sia in italiano.

Potenziali problemi di ordine (*learning effects* o deterioramento delle prestazioni causa stanchezza). Per minimizzare questa possibilità ogni traduttore segue un ordine casuale nelle traduzioni dei tre testi.

Applicazione

Collezionate due serie di dati: *on-line* e *off-line*.

Online: dati raccolti durante il processo di traduzione, log file con le battiture e le variazioni proposte da ogni soggetto.

Offline: le versioni tradotte dei tre testi originari.

Interesse qui rivolto ai dati offline. I testi di partenza segmentati secondo un'analisi lessico-grammaticale; dati raccolti su ogni segmento, sia sotto il profilo lessicale, sia sotto il profilo strutturale delle traduzioni proposte e costruita tabella relativa.

	segm1	**segm2**	**segm3**	**segm4**	**segm5**	**...**
Traduttore 1	**a**	**a**	**a**	**a**	**a**	**...**
Traduttore 2	**a**	**b**	**a**	**b**	**b**	**...**
Traduttore 3	**a**	**a**	**a**	**a**	**c**	**...**
Traduttore 4	**a**	**b**	**b**	**b**	**a**	**...**
Traduttore 5	**b**	**c**	**c**	**c**	**d**	**...**

Applicazione

Totale di 6 tabelle (3 testi * 2 tabelle a testo, lessico e strutture).

Cluster Analysis effettuate su ognuna della tabelle, per valutare la somiglianza o la dissimilarità delle prestazioni dei partecipanti, tenendo conto sia del profilo lessicale sia di quello strutturale.

Per ogni tabella effettuata l'analisi applicando i metodi del legame semplice, del legame completo, del centroide e il Metodo di Ward, valutando di volta in volta la qualità dei risultati ottenuti sulla base dei dendogrammi.

Applicazione

I risultati:

ST1	ST2	ST3
CL1:{CB,MS,MT}	CL1:{CB,MS}	CL1:{CB}
CL2:{SC, VM}	CL2:{MT,SC,VM}	CL2:{MS,VM}
		CL3:{MT,SC}

I livelli di metaforicità calcolati per i testi erano: ST3 < ST1 < ST2.

Fra ST1 e ST2 variazione nella prestazione del soggetto MT, in ST3 situazione più eterogenea, con tre cluster totali (evidenziati maggiormente sulle strutture usate più che sul lessico).

Conclusioni

I risultati dimostrano effettivamente una variazione nelle prestazioni dei partecipanti a seconda del livello di difficoltà intrinseca al testo, dovuta alla presenza più o meno marcata di *grammatical metaphor*.

Avendone la possibilità, integrare i dati di tipo *on-line* raccolti durante il processo traduttivo e i dati sugli errori commessi potrebbe rendere più efficiente l'analisi sulle prestazioni dei soggetti.